China Petroleum Processing and Petrochemical Technology

Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

BIOINFORMATICS INVESTIGATION OF GENE EXPRESSION BIOMARKERS ASSOCIATED WITH BREAST CANCER: A MACHINE LEARNING-BASED APPROACH

J. Jayapriya
Research Scholar, Dept. of Computer Science and Engineering, Annamalai University
Annamalainagar – 608002, Tamilnadu, India

Dr. G. Ramachandran
Associate Professor, Dept. of Computer Science and Engineering, Annamalai University
Annamalainagar – 608002, Tamilnadu, India

Dr.T.Priyaradhikadevi
Professor & Head, Dept. of Computer Science and Engineering, Mailam Engineering College
Mailam – 604304, Tindivanam, Tamilnadu, India

ABSTRACT
Breast cancer is the second most disease with a higher malignancy rate worldwide. The mortality level has decreased steadily over the past few decades due to various factors such as early diagnosis, medical accessibility, and effective treatment strategies. Genetic variability plays an important role in the progression of breast cancer. Analyzing the genetic profiles of the affected individual increases the chance of identifying the candidate biomarker of breast cancer. But gene expression analysis is an arduous task due to its inherent data dimensionality. Advanced computational methods act as a tool to handle complex data. This paper aims to find the candidate genetic marker associated with the condition's progression through gene expression profiling. An intelligent framework is proposed in search of biomarkers by processing the data under different levels. The dataset is accessed from the Gene Expression Omnibus repository. This paper reveals new biomarkers and novel gene pathways of breast cancer related to other diseases. The identified genetic markers are trained and validated with supervised machine-learning classification algorithms under 10-fold cross-validation. The performance of the models is evaluated with standard validation metrics. The proposed framework attained 98.76% accuracy in classifying the microarray gene expression dataset samples. The system's performance is benchmarked with other feature selection approaches, out of which the proposed framework shows better results under various validation schemes. These computational frameworks are highly beneficial for medical practitioners to diagnose individuals with more care.

INDEX TERMS Biomarker, Breast Cancer, Computational biology, Gene expression profiling, Machine learning, Microarray.

1. INTRODUCTION
Despite the rapid growth in the number of breast cancer cases, the survivability of the condition is also increasing as the treatment methodologies are highly improved by adopting advanced

Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

technologies. But, breast cancer still stands at the top, the leading cause of mortality raised due to cancer-related cases among women worldwide [1]. Many treatments, such as hormone therapy, surgery, chemotherapy, radiotherapy, etc., are suggested to effectively manage the condition [2-4]. But, the response from the treatment solely depends on the patient's condition and varies concerning multiple factors.

Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) techniques are proven to be more traditional, promising methods for accurate diagnosis [5-7]. However, it couldn't provide information related to the progression mechanism. Conversely, gene expression profiling techniques, for instance, DNA microarrays, have generated high throughput data of the genes. Gene expressions provide a snapshot of the genetic susceptibility from analyzing normal and affected cases [8].

Many statistical methods are available to make a pilot analysis of the genes. Nevertheless, finding genetic markers is still challenging due to the complex dimension of the genetic profiles [9]. Cancer occurs mostly because of genetic changes such as mutation, damage in DNA, modifications in nucleotides, and so on [10]. Therefore, tracing the abnormalities in the genes brings up a profound solution to identify the biomarkers of the disease. In this paper, the DNA microarray gene expression profiles of breast cancer-affected individuals are analyzed under various phases with statistical methods, and finding the biomarkers with effective gene selection strategy.

The rest of the article contains the following sections. The "Background study" section briefs the literature related to this study. The methodologies and framework proposed in the previous studies were reviewed with the techniques incorporated to build the model. Further, the "Materials and methods" exhibit the proposed work under detailed subsections. The results section presents the study's outcome with tables and graphical representations. The proposed study's importance and findings are discussed in the conclusion.

2. BACKGROUND STUDY

Gene expression profiles contain rich information about the characteristics, behavior, and other traits in encoded form. Analyzing these data is crucial as the data dimensions are difficult to handle. The number of samples often becomes much less with thousands of features, and this condition is also termed the "curse of dimensionality [11]". Many computational methods have recently been developed with sophisticated frameworks for complex systems, overriding conventional algorithms' difficulties. Machine learning becomes more effective in finding hidden patterns in the data generated under various circumstances [12]. Many literature studies have shown the importance of machine learning in dealing with "chaotic" systems, especially in multi-faceted healthcare modeling.

The microarray gene expressions of breast cancer are analyzed to trace the condition's progression through the modified logistic regression algorithm. The proposed methodology is tested on a series of genetic profiles, GSE25055, GSE65194, and GSE20711. Alongside, the identified biomarkers are analyzed with gene regulatory networks, and based on the observation, MCF-7 cell lines provide useful information related to the study. Also, many other biological indicators are revealed in this experiment, which shows the importance of this study [13]. The survivability of breast

Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

cancer-affected individuals is calculated using machine learning algorithms using the subtypes from the transcriptomic profiles. GSE1456 and GSE20711 datasets were considered to perform the study, which contains around 400 samples of case and control. An empirical Bayes algorithm reduced high dimensional features from 12000 to 398 features. The system's accuracy is 86%, with an AUC rate of 95% [14].

Another study employed different classifier types containing linear, non-linear, probabilistic, bagging, and boosting models and tested against the dataset with and without selecting the features. A two-layered feature selection was performed, where 50 candidates were identified. Then the performance is calculated on these features, and SVM performs better than the rest of the models [15]. This study adopts a machine learning-related approach to find the genetic markers to guide the proper treatment of breast cancer-affected individuals. Three hundred forty-seven samples were collected to conduct this experiment, with 4066 features identified after feature selection from the original dimension. This system deals with a flow of phases, starting with class imbalance, followed by feature selection where mRMR is used and ends up with a multiclass classification algorithm, especially SVM, RF, and NB. This study revealed several biomarkers associated with the disease through ML approaches [16].

3. MATERIALS AND METHODS

This section briefs about the proposed experimental work in different phases. The microarray gene expression dataset is discussed with its properties, followed by the processing methods of raw data, transforming into a cell of probes. Then the selection of candidate features is performed in consecutive pipelines, and the biomarkers are identified through the DEG-mRMR model. The performances of the findings are calculated through machine learning algorithms under the standard performance evaluation metrics.

3.1 DATASET INFORMATION

The microarray dataset is accessed from the Gene Expression Omnibus (GEO) repository [17, 18], managed by the National Center for Biotechnology Information (NCBI) [19]. The accession number is GSE139038 [20]. The dataset's raw form (. CEL) is fetched to conduct the study. A detailed description of the dataset is given in Table 1.

table 1 dataset description

Details	Source Information
Data Repository	Gene Expression Omnibus
Accession Number	GSE109169
Disease Type	Breast Cancer
Type of Data	DNA Microarray
Number of Samples	50
Number of Features	19076
Case (Disease)	25 Samples

Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

Control (Adjacent Tissue)	25 Samples
Data Type	Numeric (Continuous)

3.2 DATA PREPROCESSING

The. CEL formatted file is initially processed through the "limma" library [21] supported by the "Bioconductor" [22] package in the R language. The "read. maimages" function imports the dataset into the playground. Meanwhile, the background correction [23] uses the "normexp" function. The quantile normalization [24] method is applied to perform adjustments in the transformed data. The resultant dataset consists of the probes as features in the dataset. The complete architecture of the proposed system is given in Figure 1.

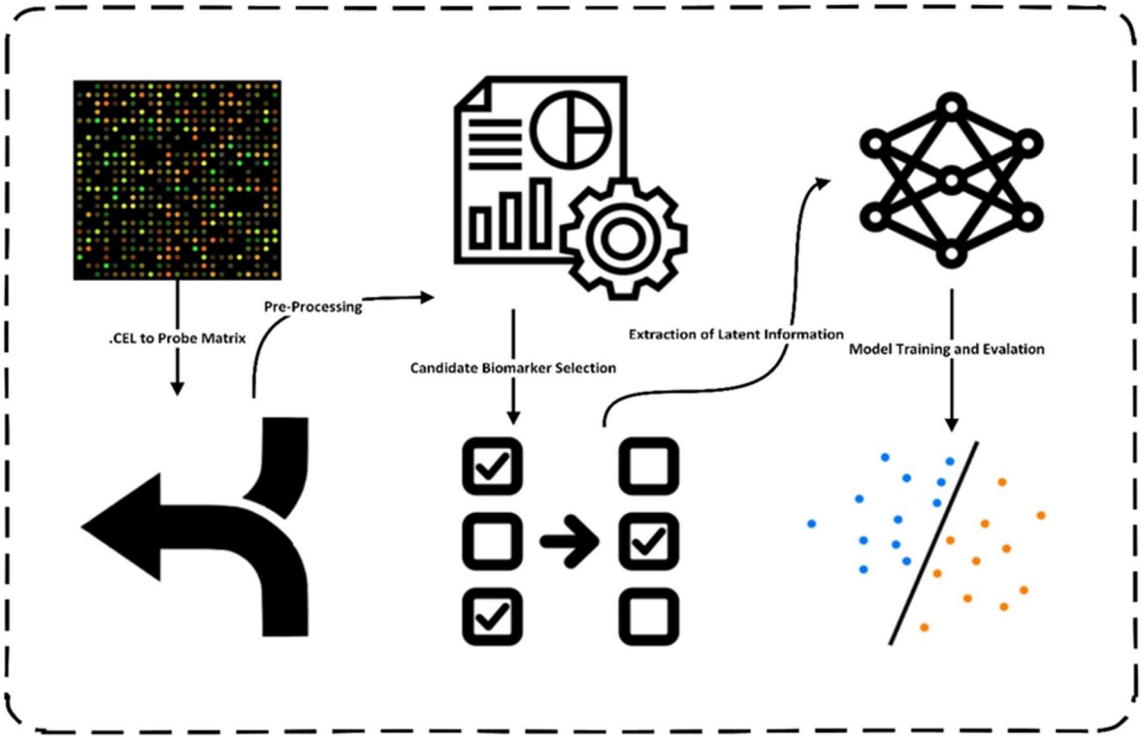


Figure 1 Processing of Information in the Proposed Pipeline

3.3 INFORMATIVE GENE SELECTION

The high-dimensional gene expression dataset has many gene probes, and not all are important in predicting the condition. So, selecting the informative genes has a greater influence on improving model performance. Here, the t-test is implemented to find the Differentially Expressed Genes (DEGs) that have shown differences between samples of subgroups [25-27]. The top 250 genes were initially filtered from the entire set using the "GEO2R" [28] library. To avoid multiple testing errors in this process, the moderated t-test is performed by adjusting the p-values with Bonferroni correction [29].

3.4 BIOMARKER IDENTIFICATION

Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

In the previous phase, the significant DEGs are identified with a t-test where the adjusted p-value is less than 0.05. Not all of these top 250 features are said to be the biomarker of breast cancer. So, finding the optimal candidate markers from the 250 selected features is performed with minimum redundancy and maximum relevance (mRMR) [30].

The mutual information for a set of features X and Y is represented in eqn. 1

$$M(X,Y) = \sum_{c\in C} \sum_{b\in B} p\,(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

The maximum relevance between the feature and target class is represented in eqn. 2

$$max\ Rl(F,t); D = \frac{1}{|F|}\sum_{F_i\in F} M\,(F_i;t) \quad (2)$$

Similarly, minimum redundancy between each feature against the rest of the feature can be calculated using eqn. 3

$$minRd(F); Rd = \frac{1}{|F|}\sum_{F_i,F_j\in F} M\,(F_i,F_j) \quad (3)$$

The final feature subset S is identified based on the scores attained from eqn. 2 and 3 through eqn. 4 is written as

$$mRMR = max_{F_i\notin S}[M(F_i;t) - \frac{1}{|S|}\sum_{F_j\in S} M\,(F_j;F_i)] \quad (4)$$

This method finds 21 discriminative biomarkers by eliminating the redundant features and holding the features with higher relevancy with the target class. The dataset with these features is trained with machine learning classifiers. The performance sees a slight hike over the 250 markers, and the current result is optimal.

TABLE 2 NUMBER OF FEATURES SELECTED IN EACH PHASE

Initial Features	the t-test (DEGs)	DEG-mRMR
19076	250	21

3.5 PREDICTIVE MODELING WITH SUPERVISED CLASSIFICATION ALGORITHMS

The feature subsets 250, 21 and the compressed vectors are trained with supervised classifiers Random Forest (RF), Linear Discriminant Analysis (LDA), Linear Support Vector Machine (SVM), and Naïve Bayes (NB) [31-33]. The training and validation data set is generated by a 10-fold cross-validation method [34, 35]. Table 2 lists the number of features selected in each phase.

4. RESULTS

This experimental study aims to find the biological signatures of breast cancer from the microarray gene expression dataset. This exploratory model identified 21 biomarkers from the gene expression probes. The features are then inputted into the classifiers to calculate the performance to portray the efficacy of the study. Additionally, Correlation-based Feature Selection (CBFS) [36], Bat Search Optimization (BSO) [37], and Conditional Mutual Information Maximization (CMIM) [38] techniques are deployed to benchmark the performance against the proposed feature selection method.

Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

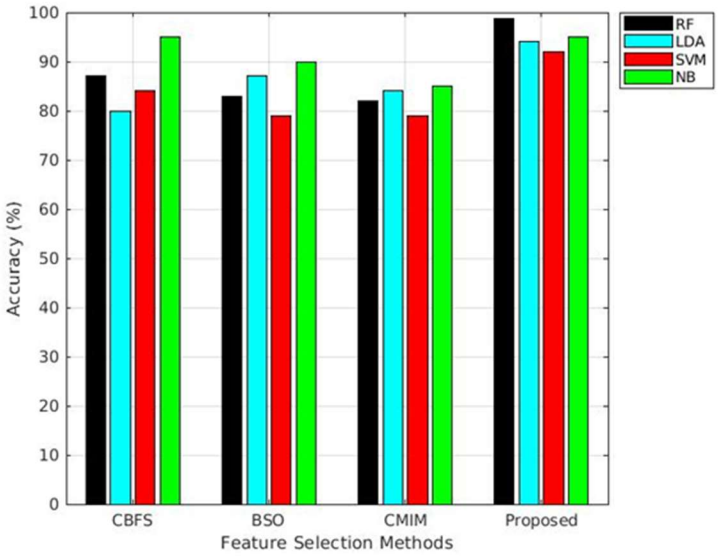


Figure 2 Comparison of performance between different feature selection methods with the proposed model

The results obtained from all the methods are compared, and the best performance is achieved from the DEG-mRMR model. The scores are calculated for accuracy, false-positive rate, sensitivity, specificity, and f-score. In Table 3, the scores of the models are given. Among all, DEG-mRMR features trained by RF show better results than the rest of the classifiers by attaining an accuracy of 98.76%. Figure 2 and 3 exhibit performance under different measures.

TABLE 3 PERFORMANCE OF CLASSIFIERS ON DIFFERENT FEATURE SELECTION TECHNIQUES IN (%)

GSE109169	Metrics	RF	LDA	SVM	NB
CBFS	**Acc**	94.56	91.94	93.07	91.65
	FPR	8.24	07.86	08.54	08.40
BSO	**Acc**	89.94	87.26	87.14	84.42
	FPR	6.89	9.28	10.51	9.23
CMIM	**Acc**	91.86	90.41	86.79	88.46
	FPR	6.21	7.41	8.67	9.26
Proposed	**Acc**	98.76	94.62	92.54	94.19
	FPR	1.68	4.56	6.84	5.39

Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

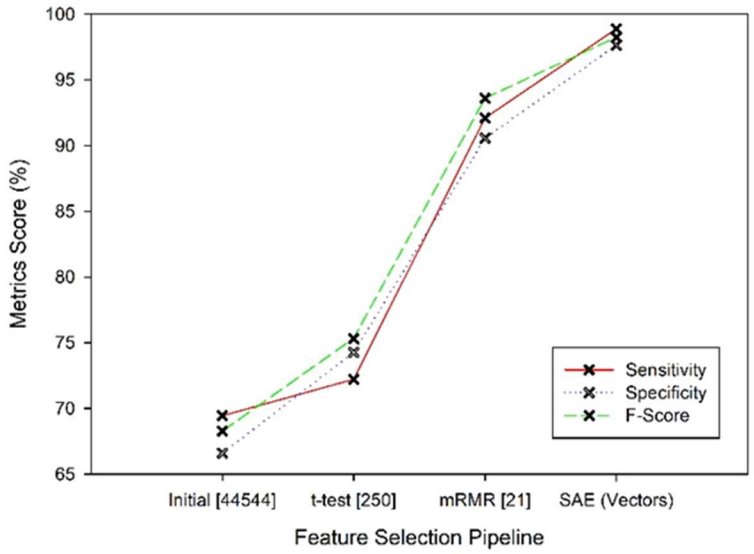
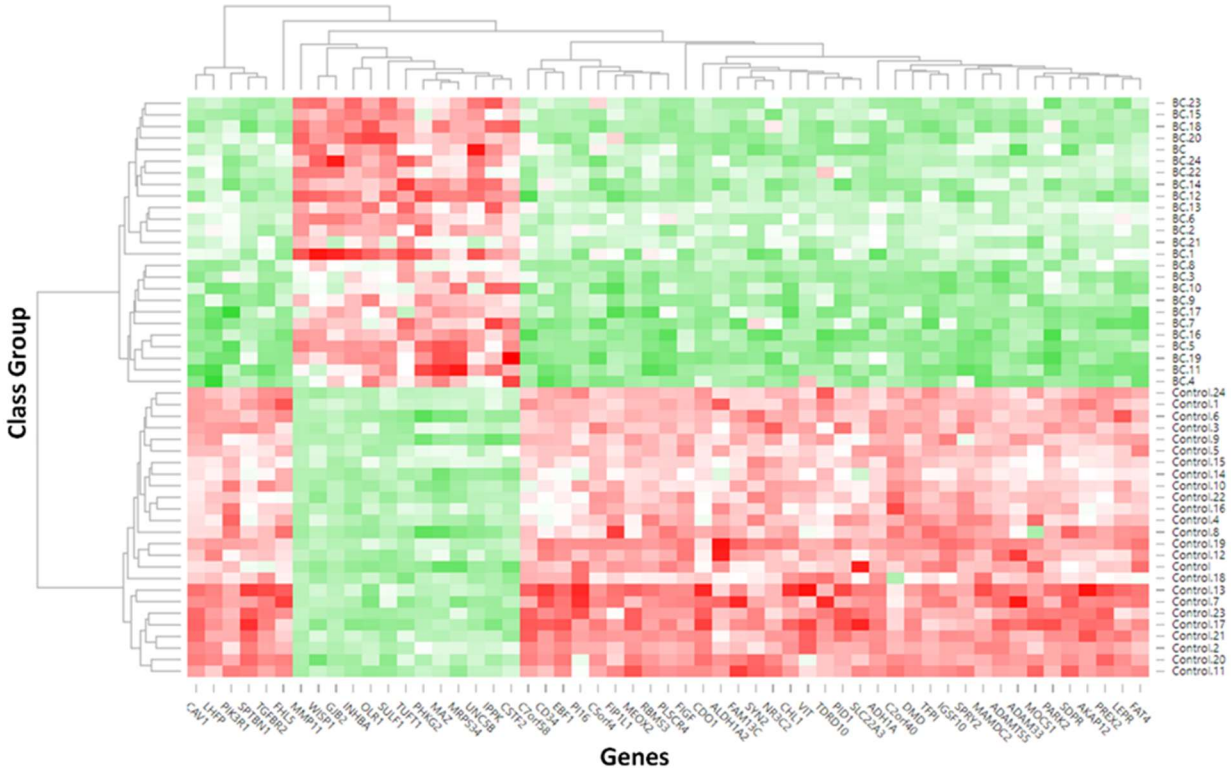


Figure 3 Performance of the features from the proposed system in each phase with RF Classifier

4.1 HEATMAP ANALYSIS

A system of color codes representing the range of values in a different form in a graphical way is said to be a heat map. In microarray gene expression data, the value distribution is continuous. The feature subset generates the heat map to find the relationship between the case and control samples [39]. Also, the heat map gives more visibility and a comprehensive view of the high-dimensional data. Figure 4 shows the projected heat map representation of the identified 21 biomarkers.



Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

Figure 4 Heatmap of the Biomarkers identified through mRMR

4.2 GENE SET ENRICHMENT ANALYSIS

Gene Set Enrichment Analysis (GSEA) is a bioinformatics and genomics computational approach for analyzing gene expression data and determining if preset collections of genes, known as gene sets or gene ontologies, are significantly enriched in a particular dataset. GSEA is very beneficial for detecting functional connections and biological pathways expressed differently in different experimental circumstances or phenotypes. GSEA works on ranking all genes in a dataset based on their differential expression levels between two or more conditions.

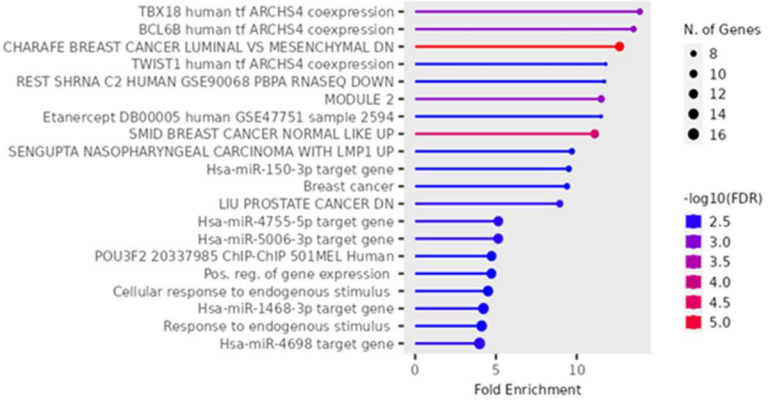
GSEA gene sets can be obtained from various sources, including functional annotations, recognized pathways, or gene signatures linked to certain biological processes or disorders. The gene enrichment is calculated using the shinyGO webserver. Table 4 provides enrichment results based on the false discovery rate (FDR), the number of genes involved, and their pathways. The outcome reveals the interaction of the biomarkers with breast cancer. Figure 5 illustrates the barplot and network plot representation of the pathways associated with gene enrichment on the biomarker genes.

TABLE 4 Gene enrichment scores and their pathways

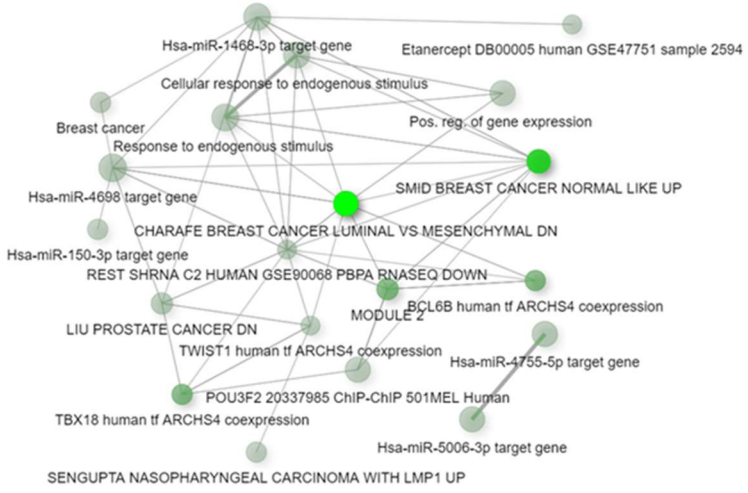
Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways
7.7E-04	8	279	13.9	TBX18 human tf ARCHS4 coexpression
7.7E-04	8	287	13.5	BCL6B human tf ARCHS4 coexpression
3.6E-06	12	460	12.7	CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL DN
3.5E-03	7	288	11.8	TWIST1 human tf ARCHS4 coexpression
3.5E-03	7	290	11.7	REST SHRNA C2 HUMAN GSE90068 PBPA RNASEQ DOWN
7.5E-04	9	379	11.5	MODULE 2
3.6E-03	7	295	11.5	Etanercept DB00005 human GSE47751 sample 2594
4.6E-05	11	480	11.1	SMID BREAST CANCER NORMAL LIKE UP
3.5E-03	8	400	9.7	SENGUPTA NASOPHARYNGEAL CARCINOMA WITH LMP1 UP

3.5E-03	8	408	9.5	Hsa-miR-150-3p target gene
3.5E-03	8	413	9.4	Breast cancer
3.1E-03	9	488	8.9	LIU PROSTATE CANCER DN
3.1E-03	13	1225	5.1	Hsa-miR-4755-5p target gene
3.1E-03	13	1227	5.1	Hsa-miR-5006-3p target gene
3.5E-03	13	1334	4.7	POU3F2 20337985 ChIP-ChIP 501MEL Human
3.5E-03	13	1336	4.7	Pos. reg. of gene expression
3.5E-03	14	1505	4.5	Cellular response to endogenous stimulus
3.4E-03	15	1723	4.2	Hsa-miR-1468-3p target gene
3.5E-03	15	1769	4.1	Response to endogenous stimulus
3.2E-03	16	1948	4	Hsa-miR-4698 target gene

(A)



(B)



Catalyst Research Volume 23, Issue 2, December 2023 Pp. 4652-4663

Figure 5 Enrichment Network Association of Biomarker genes (A) Barplot, (B) Network plot

5. CONCLUSION

This paper proposes an effective breast cancer classification model from the high-dimensional microarray gene expression profiles. In the initial phase, the raw format of data from the sequencer is converted into an array of probe values with the RMA technique. The transformed data contains hundreds and thousands of genetic features identified as probe values. This represents the features of the samples considered for predicting samples into various categories. The complexity of the model becomes very high if all the features are used to train the model. Moreover, not all features contain useful information for accurate predictions through effective discrimination. So, in the next phase, genetic feature selection is performed based on p-value with the significance of $p<0.05$. The features selected under these criteria are considered informative. Then, the predictor features are identified with the minimum redundancy maximum relevance algorithm. The performance of the features selected on the mRMR feature selection method is evaluated with supervised classification algorithms. The bioinformatics analysis, such as heatmap modeling and gene set enrichment analysis, reveals the significant association between the genes identified in the present study with breast cancer. Also, the performance is benchmarked with existing feature selection methods, where the proposed DEG-mRMR algorithm exhibits better results with 98.76% accuracy. This model performance will be further improved by implementing advanced autoencoder models with less complexity.

REFERENCES

- Shapiro, S., Venet, W., Strax, P., Venet, L., & Roeser, R. (1982). Ten-to fourteen-year effect of screening on breast cancer mortality. Journal of the National Cancer Institute, 69(2), 349-355.
- Li, C. I., Malone, K. E., Porter, P. L., Weiss, N. S., Tang, M. T. C., Cushing-Haugen, K. L., & Daling, J. R. (2003). Relationship between long durations and different regimens of hormone therapy and risk of breast cancer. Jama, 289(24), 3254-3263.
- Darby, S. C., Ewertz, M., McGale, P., Bennet, A. M., Blom-Goldman, U., Brønnum, D., ... & Jensen, M. B. (2013). Risk of ischemic heart disease in women after radiotherapy for breast cancer. New England Journal of Medicine, 368(11), 987-998.
- Gärtner, R., Jensen, M. B., Nielsen, J., Ewertz, M., Kroman, N., & Kehlet, H. (2009). Prevalence of and factors associated with persistent pain following breast cancer surgery. Jama, 302(18), 1985-1992.
- Balu-Maestro, C., Chapellier, C., Bleuse, A., Chanalet, I., Chauvel, C., & Largillier, R. (2002). Imaging in evaluation of response to neoadjuvant breast cancer treatment benefits of MRI. Breast cancer research and treatment, 72(2), 145-152.
- Pace, L., Nicolai, E., Luongo, A., Aiello, M., Catalano, O. A., Soricelli, A., & Salvatore, M. (2014). Comparison of whole-body PET/CT and PET/MRI in breast cancer patients: lesion detection and quantitation of 18F-deoxyglucose uptake in lesions and in normal organ tissues. European journal of radiology, 83(2), 289-296.

- Yang, W. T., Le-Petross, H. T., Macapinlac, H., Carkaci, S., Gonzalez-Angulo, A. M., Dawood, S., ... & Cristofanilli, M. (2008). Inflammatory breast cancer: PET/CT, MRI, mammography, and sonography findings. Breast cancer research and treatment, 109(3), 417-426.
- Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., ... & Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proceedings of the National Academy of Sciences, 99(20), 12963-12968.
- Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., & Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. The American Journal of Human Genetics, 69(1), 138-147.
- Ford, D., Easton, D. F., Bishop, D. T., Narod, S. A., & Goldgar, D. E. (1994). Risks of cancer in BRCA1-mutation carriers. The Lancet, 343(8899), 692-695.
- Catchpoole, D., Kennedy, P., Skillicorn, D., & Simoff, S. (2010). The curse of dimensionality: a blessing to personalized medicine. Verified OK.
- Anzai, Y. (2012). Pattern recognition and machine learning. Elsevier.
- Morais-Rodrigues, F., Silvério-Machado, R., Kato, R. B., Rodrigues, D. L. N., Valdez-Baez, J., Fonseca, V., ... & Dutra, J. D. C. F. (2020). Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. Gene, 726, 144168.
- López-González, K., & Dávila, C. (2017). Predicting Survivability using Breast Cancer Subtype with Transcriptomic Profiles. In IIE Annual Conference. Proceedings (pp. 1406-1411). Institute of Industrial and Systems Engineers (IISE).
- Turgut, S., Dağtekin, M., & Ensari, T. (2018, April). Microarray breast cancer data classification using machine learning methods. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-3). IEEE.
- Abou Tabl, A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Frontiers in Genetics, 10, 256.
- Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207-210.
- Davis, S., & Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics, 23(14), 1846-1847.
- Coordinators, N. R. (2017). Database resources of the national center for biotechnology information. Nucleic acids research, 45(Database issue), D12.
- https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE109169 (Accessed on 24-07-2023)
- Smyth, G. K., Ritchie, M., Thorne, N., & Wettenhall, J. (2005). LIMMA: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health.

- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... & Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology, 5(10), R80.
- Harbron, C., Chang, K. M., & South, M. C. (2007). RefPlus: an R package extending the RMA Algorithm. Bioinformatics, 23(18), 2493-2494.
- López-Romero, P., González, M. A., Callejas, S., Dopazo, A., & Irizarry, R. A. (2010). Processing of Agilent microRNA array data. BMC research notes, 3(1), 18.
- Wright, G. W., & Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics, 19(18), 2448-2455.
- Hatfield, G. W., Hung, S. P., & Baldi, P. (2003). Differential analysis of DNA microarray gene expression data. Molecular microbiology, 47(4), 871-877.
- Cui, X., Hwang, J. G., Qiu, J., Blades, N. J., & Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics, 6(1), 59-75.
- Karthik, S., & Sudha, M. (2020). Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network. Evolutionary Intelligence, 1-16.
- Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. Bmj, 310(6973), 170.
- Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC bioinformatics, 18(1), 9.
- Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.
- Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7), 1895-1923.
- Aly, M. (2005). Survey on multiclass classification methods. Neural Netw, 19,1-9.
- Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48(9), 2839-2846.
- Koyejo, O. O., Natarajan, N., Ravikumar, P. K., & Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. In Advances in Neural Information Processing Systems (pp. 2744-2752).
- Hall, M. A. (1999). Correlation-based feature selection for machine learning.
- Wang, G., & Guo, L. (2013). A novel hybrid bat algorithm with harmony search for global numerical optimization. Journal of Applied Mathematics, 2013.
- Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. Journal of Machine learning research, 5(Nov), 1531-1555.
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849.